第 38 卷 第 6 期
2023 年 12 月

天津科技大学学报
Journal of Tianjin University of Science & Technology

Vol. 38 No. 6
Dec. 2023

DOI:10.13364/j.issn.1672-6510.20230016

知识嵌入的医疗对话生成

王 嫄，曾磊磊，武振华，熊 宁
（天津科技大学人工智能学院，天津 300457）

摘 要：针对以往医疗对话生成方法未能有效建模医学知识，导致生成的回复缺乏医学常识一贯性的问题，很多学者尝试引入医疗知识图谱，但集成医疗知识图谱时容易占用较多输入数据空间，这限制了模型输入可以保留对话上下文信息量的大小。本文提出知识嵌入的医疗对话生成模型（medical conversation generation model based on knowledge embedding，MCG-KE），该模型基于历史对话进行实体预测得到上下文知识嵌入实体，引入串行图编码方式和图注意力机制获得当前对话相关的医疗知识图谱子图编码，将上下文知识嵌入实体、医疗知识图谱子图编码和历史对话编码作为对话生成模型的输入，用于知识嵌入的医疗对话生成。实验结果表明，模型在高效计算的情况下，所生成的医疗对话在自动评价和人工评价等相关指标上的性能均有提升。

关键词：医疗对话生成；上下文知识嵌入；知识图谱子图编码；图注意力机制；生成模型
中图分类号：TP391 **文献标志码**：A **文章编号**：1672-6510（2023）06-0054-08

Knowledge-Embedded Medical Dialogue Generation

WANG Yuan，ZENG Leilei，WU Zhenhua，XIONG Ning
（College of Artificial Intelligence，Tianjin University of Science & Technology，Tianjin 300457，China）

Abstract：Aiming at the problem of lack of medical common senses consistency caused by the failure of medical knowledge modeling in previous medical conversation generation methods，many researchers have tried to introduce the medical knowledge graph，but it was easy to occupy too much input data space when integrating the medical knowledge graph，which has limited the dialogue context information that could be retained by the model input. In this article，a medical conversation generation model based on knowledge embedding（MCG-KE） is proposed. This model makes entity prediction based on historical dialogue to obtain context knowledge embedded entity，and introduces serial graph coding and graph attention mechanism to obtain the subgraph coding of medical knowledge graph related to current dialogue. Context knowledge embedding entity，medical knowledge spectrum subgraph coding and historical dialogue coding are used as input of dialogue generation model for knowledge embedding medical conversation generation. The experimental results showed that，under the condition of efficient calculation，the performance of the medical conversation generated by the model was improved in the relevant indexes such as automatic evaluation and manual evaluation.

Key words：medical conversation generation；contextual knowledge embedding；knowledge graph spectrum subgraph coding；graph attention mechanism；generation model

现如今，医疗资源紧张的情况普遍存在，根据2020 年中国卫生健康统计年鉴[1]，全国平均每千人的医疗机构床位数为 6.30 张，其中：东部地区平均每千人的医疗机构床位数为 5.78 张，中部地区平均每千人的医疗机构床位数为 6.44 张，西部地区平均每千人的医疗机构床位数为 6.84 张。不同地域情况不同，西部地区地广人稀，千人床位数相对较高，而东部地区人口密度较大，千人床位数相对较低。倘若人人都去线下医院咨询，这势必会给医院造成很大的压力，并且容易引起交叉感染，导致更加严重的后果。因

收稿日期：2023-02-04；**修回日期**：2023-04-18
基金项目：国家自然科学基金项目（61976156）；大学生创新创业训练计划项目（202210057063）
作者简介：王 嫄（1989—），女，山西太原人，副教授，wangyuan23@tust.edu.cn

此，一个可以让患者在线咨询和问诊的医疗对话系统就显得十分必要。它能够节省人力、物力，减小线下医院的压力，提高问诊效率，把有限的医疗资源留给最需要的群体，更好、更及时地满足患者的需求。

知识图谱本质上是语义网络的知识库，它主要包含实体、属性和关系。实体是指现实世界中的事物，比如人、地名、概念、药物、公司等；属性是指实体具备的某些特征，属性值即为某种属性相对应的具体值；关系则用来表达不同实体之间的某种联系。现有基于知识图谱的医疗对话生成的基本思想是在生成回复时向模型提供所需要的实体信息和关系信息。然而，回复生成模型的输入序列长度通常是有限的。例如，对于 GPT-2（generative pretrained transformer，GPT）模型来说，输入序列长度不能超过 1 024 个词，因而背景知识的引入会高度影响可以输入模型的上下文信息量的大小。在早期工作尝试的方法中，图谱中的局部知识被改写成伪话语并与对话历史中的话语一起提供给模型，但这种做法容易丢失知识图谱中的实体关系逻辑推理信息。

在医疗对话生成中，为了提高模型生成回复的内容准确性和对话信息量，很多学者[2-3]尝试根据历史对话引入相关疾病实体信息，但是这些方法在多轮的长对话中仍然存在话语之间联系不紧密、医疗信息表述多样化、医疗信息利用不足的问题。

为了解决上述问题，本文提出知识嵌入的医疗对话生成模型（medical conversation generation model based on knowledge embedding，MCG-KE），根据历史对话进行实体预测得到上下文知识嵌入实体，引入串行图编码方式和图注意力机制获得当前对话相关的医疗知识图谱子图编码，并保留子图在其结构中的编码信息。该模型通过实体预测模块得到上下文知识嵌入实体，利用历史对话中的实体生成知识图谱子图，同时获取子图的语义信息，利用图注意力机制表示实体之间的关系，使用串行图编码技术简洁地编码知识图谱子图，将上下文知识嵌入实体、历史对话和子图编码结果作为模型的输入，从而使生成的回复具有医学常识一贯性。

1　相关工作

在对话系统中产生基于知识的回复是一个重要的研究挑战。知识图谱可以被视为现实世界的一个抽象概念，它可以潜在地促进对话系统产生基于知识的回复。然而，以端到端方式将知识图谱集成到对话生成过程中是一项艰巨的任务。Chaudhuri 等[4]将知识图谱集成到回复生成过程中，训练一个 BERT（bidirectional encoder representations from transformers）模型，学习在多任务的端到端设置中使用知识图谱的元素进行回答。在使用图拉普拉斯函数进行训练和推理的过程中，将知识图谱的 k 跳子图纳入模型中。引入外部知识，通过选择具体的内容加到回复生成过程中，提升回复的质量。但知识为实体，无法为回复生成提供其他更加丰富的信息，而且非结构化的表示方案要求模型具有很强的能力从知识文本集合中进行知识选择。Liu 等[5]提出基于扩充知识图的开放域对话生成模型（AKGCM），融合非结构化知识和结构化知识，模型由知识选择和回复生成这两个模块组成。知识选择模块转化为一个多跳图问题，基于强化学习的推理模型（MINERVA）有效捕获会话流，实现知识选择，回复生成模块使用带复制机制的编码器、解码器模型，基于所选知识和用户输入生成最终回复。通过引入外部背景知识，神经对话模型可以在生成流畅和信息丰富的回复方面显示出巨大的潜力。然而，构建这种以知识为基础的对话很费力，而且现有模型在迁移到训练样本有限的新领域时通常表现不佳。基于弱监督学习的新型三阶段学习框架（TSLF）受益于大规模的对话和非结构化知识库，同时作者还设计了带有解耦解码器的 Transformer 变体，促进了响应生成和知识整合的分离学习[6]。

本文研究基于知识的医疗对话生成。知识的引入可以丰富模型生成回复的信息量，缓解高频无实际意义的回复的问题。目前，基于知识的医疗对话生成主要分为基于实体知识的医疗对话生成和基于图谱知识的医疗对话生成。

基于实体知识的医疗对话生成主要通过挖掘与上下文相关的医学实体，辅助回复生成。研究人员分别使用检索方法和生成方法，将医学实体知识用于医疗对话生成任务，其中检索方法通过使用医学实体作为关键信息在语料库中检索与该医学实体最相关的回复，而生成方法则将医疗实体编码并作为序列到序列模型的输入逐词生成回复[2]。

基于图谱知识的医疗对话生成包括引入图谱推理技术的医疗对话生成和引入图谱融合技术的医疗对话生成，主要利用医疗知识图谱实现。

引入图谱推理技术的医疗对话生成工作如下。通过构造一种全局注意力机制以及症状图模拟症状

之间的关联，提高了在医疗对话中对于每一句话出现相关症状的预测准确率，以及症状推理（患者是否有某一种症状）的精度和症状诊断的性能[7]。为了解决训练数据匮乏疾病的医疗对话生成模型学习困难的问题，Lin 等[8]提出低资源医疗回复生成器，该模型集成了分层上下文编码器、元知识图推理网络和图指导的响应生成器，利用可动态进化的常识图反映和推断疾病与症状的相关性，将问诊模式从数据资源丰富疾病迁移到数据资源匮乏疾病。

引入图谱融合技术的医疗对话生成工作如下。为了全覆盖复杂的医疗对话自然语言模式和场景，Xu 等[9]提出了知识路由关系对话系统，该系统将丰富的医疗知识图谱融入对话管理的主题切换中，并且使其与自然语言理解和自然语言生成协作。系统使用知识路由 DQN（KR-DQN）管理主题切换，它集成了一个关系细化分支编码不同症状和症状疾病对之间的关系，以及一个用于主题决策的知识路由图谱分支。为了使问答型医疗对话系统检索出来的答案更加科学合理，章毅等[10]提出了一种基于语言模型和实体匹配的问答型医疗对话系统构建方法。作者收集网络医疗讨论帖清洗后存入 ElasticSearch 中作为检索数据集；使用医疗自然语言处理比赛数据集的开源数据，训练出医疗相关的命名实体识别模型；收集开源网站的公开数据集构成医疗知识图谱，扩充检索流程。作者构建的基于语言模型和实体匹配的问答型医疗对话系统在经过召回、精排和综合评分几个步骤之后，结合合理的评分机制，输出一个最为合适的回答，弥补检索式问答系统和知识图谱式问答系统的缺陷。穆天杨等[11]将知识图谱中的结构化信息应用到对话系统中，提出了融合医疗知识图谱的端到端对话系统。使用词表匹配与深度学习方法相结合的方式提取对话信息中的关键词，并根据提取的关键词搜索知识图谱中的相关信息作为 GPT-2 模型的输入，从而提高端到端对话系统的表现效果。

上述基于实体知识和图谱知识的医疗对话生成方法都在对话生成过程中引入了知识，但知识之间的复杂语义线索关系仍未被有效利用，图谱的引入仍然占用较多数据空间。

2　模型与方法

2.1　问题形式化描述

知识嵌入的医疗对话生成任务定义如下。给定医生和患者历史对话的文本序列 $S=\{S_1,\cdots,S_i,\cdots,S_k\}$ 和医疗知识图谱数据集 MKG，其中 k 表示当前对话的轮次，S_i 为来自医生 d 或患者 p 的由 l_i 个词组成的话语 $S_i=\{w_{i1},\cdots,w_{il_i}\}$。任务基于历史对话 S 生成上下文一致且有医学意义的长度为 r 的文本序列 $R=\{w_{i1},\cdots,w_{ir}\}$ 作为回复为患者提供诊断建议。

2.2　模型框架

本文模型利用历史对话和医疗知识图谱引导对话生成预问诊回复，模型包含上下文知识嵌入模块、子图编码模块和回复生成模块。上下文知识嵌入模块利用数据集中的候选实体集合计算出与历史对话最相关的医学实体；子图编码模块旨在通过串行图编码技术编码从医疗知识图谱数据集中获取的子图；回复生成模型则使用上下文知识嵌入实体、历史对话和编码后的医疗知识图谱子图生成回复。首先，根据患者和医生的历史对话以及候选实体集合计算出与历史对话最相关的上下文知识嵌入实体；然后，基于历史对话数据中的医学实体从医疗知识图谱中获取子图，并使用串行图编码方法对其进行简洁编码，从而获取子图的相关语义信息；最后，将上下文知识嵌入实体、历史对话和编码的子图连接并输入 GPT-2 模型中解码生成对话回复。训练过程中，通过最小化下一个标记预测的负对数似然优化来自 GPT-2 模型的权值。基于知识嵌入的医疗对话生成模型如图 1 所示。

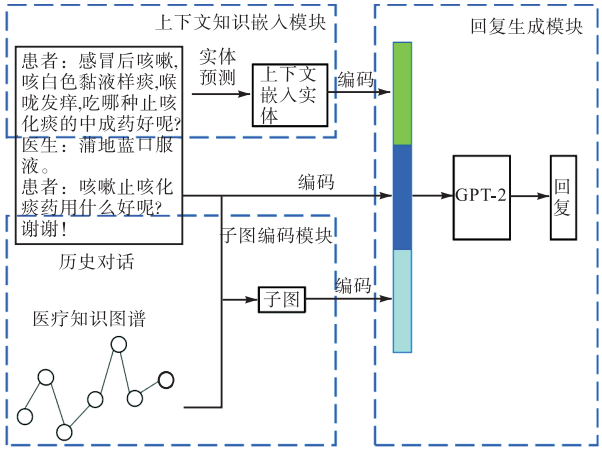

图 1　基于知识嵌入的医疗对话生成模型

Fig. 1　Medical conversation generation model based on knowledge embedding

2.3　上下文知识嵌入模块

在上下文知识嵌入模块中，首先基于医生和患者的历史对话的文本序列 S 得到对应的类型向量 $\boldsymbol{T}$ 和

位置向量 $\boldsymbol{P}$。类型向量 $\boldsymbol{T}$ 是指医生 d 或患者 p，位置向量 $\boldsymbol{P}$ 通过学习得到。将文本序列 S、类型向量 $\boldsymbol{T}$ 和位置向量 $\boldsymbol{P}$ 作为 BERT[12]的输入进行编码即得到历史对话向量 $\boldsymbol{M}_u$。

$$\boldsymbol{M}_u = \mathrm{BERT}(S + \boldsymbol{T} + \boldsymbol{P}) \tag{1}$$

为进一步提升模型对非线性复杂上下文语义的表征建模能力，本文引入由多个节点层组成，每一层全连接到下一层的多层感知机（multilayer perceptron，MLP）映射历史对话向量 $\boldsymbol{M}_u$ 到患者当次就诊上下文表示向量 $\boldsymbol{V}_p$，即

$$\boldsymbol{V}_p = \mathrm{MLP}(\boldsymbol{M}_u) \tag{2}$$

计算 $\boldsymbol{V}_p$ 和每个候选实体向量的乘积，使用 softmax 函数从数据集的候选实体集合中找到概率值最大即最相关的实体。第 i 个候选实体与当前就诊历史对话相关性概率为

$$P(y=i) = \frac{\boldsymbol{e}_i \boldsymbol{V}_p}{\sum_{i=1}^{K} \boldsymbol{e}_i \boldsymbol{V}_p} \tag{3}$$

其中 $\boldsymbol{e}_i$ 是实体 i 的词向量。

选择相关性概率最大的实体作为被预测的实体，即关键指导实体，为

$$i = \arg\max(P(y=i)) \tag{4}$$

2.4　子图编码模块

利用历史对话数据中的医学疾病实体获取医疗知识图谱中以该疾病实体为中心的子图。串行图编码见表 1。在串行图编码中，子图中的关系和实体标记在编码结果 GE 中的 words 层中串行排列，在编码结果中的 segments 层使用两个新的标记 entity 和 relation 区分 words 层的关系和实体。由于输入上下文的长度固定，通过使用串行图编码技术，模型编码知识图谱所需要的空间减少，实现了知识图谱高效编码，节省了输入数据空间。

表 1　串行图编码

Tab. 1　Serial graph coding

words 层	BOS	腹泻	symptom	消化内科	category
segments 层	BOS	entity	relation	entity	relation

为了保存子图的结构信息，本文创建并添加了子图的图注意力矩阵，矩阵中的权值设置为 1 或者 0，即当两个实体在医疗知识图谱中相邻并有边连接，矩阵的权值设置为 1，否则权值设置为 0。例如：实体吸烟中毒症、肺部检查、腹痛、高血压、布美他尼片、消化内科对应的图注意力机制矩阵如图 2 所示。

对于序列 S，GPT-2 模型在进行预测时，模型第 l 层第 i 个词的隐状态 $\boldsymbol{h}_i^l$ 计算公式如下，其中 Q、K、V 是可学习的参数，$\boldsymbol{P}_j$ 即图注意力矩阵。

$$a_{ij} = \mathrm{softmax}_j(\boldsymbol{P}_j + Q^{l-1}\boldsymbol{h}_i^{l-1} K^{l-1}\boldsymbol{h}_j^{l-1}) \tag{5}$$

$$\boldsymbol{h}_i^l = \sum_{j \in S} a_{ij}(V^{l-1}\boldsymbol{h}_j^{l-1}) \tag{6}$$

	吸烟中毒症	肺部检查	腹痛	高血压	布美他尼片	消化内科
吸烟中毒症	1	1	1	1	1	0
肺部检查	1	1	0	0	0	0
腹痛	1	0	1	0	0	0
高血压	1	0	0	1	0	0
布美他尼片	1	0	0	0	1	0
消化内科	0	0	0	0	0	1

图 2　图注意力机制矩阵

Fig. 2　Graph attention mechanism matrix

2.5　回复生成模块

回复生成模块旨在使用 GPT-2 模型生成医疗问诊回复。为了使生成的回复更加具有医疗场景语义信息，将使用子图编码模块得到的编码结果、嵌入实体 i 和历史对话共同作为 GPT-2 模型的输入，并解码生成回复

$$R = \text{GPT-2}(i, S, \mathrm{GE}) \tag{7}$$

其中：R 为模型生成的回复，GE 为子图编码的结果，S 为历史对话序列，i 为上下文知识嵌入实体。

3　实　验

3.1　实验设置

3.1.1　数据集

为了证明本方法的有效性以及鲁棒性，使用两个公开数据集 MedDG 和 CovidDialog-Chinese 以及知识图谱 MedicaKnowledgeGraph 进行实验。

MedDG 数据集[2]是一个与 12 种常见胃肠道疾病相关的大规模、高质量的医学对话数据集，它包含从中国健康咨询平台收集的超过 17 000 段对话，是目前医疗对话回复生成任务测评中的公认数据集。该数据集平均每个话语包含词数是 17.7。候选实体集合有 5 类共 160 个实体，类别包括疾病、症状、属性、检查和药物。疾病实体共 12 项，包括胃炎、肠炎、便秘等；症状实体共 62 项，包括腹泻、腹痛、腹胀等；属性实体共 4 项，包括时长、诱因、性质、位置；检查实体共 20 项，包括胃镜、肠镜、便常规等；药物实体共 62 项，包括奥美拉唑、吗丁啉、莫沙必利等。本文将该数据集划分为训练集（14 864 段对话）、验证集（2 000 段对话）和测试集（1 000 段对话）。

CovidDialog-Chinese 数据集[13]来源于 haodf.com

在线医疗服务平台，它包含 1 088 段关于新冠肺炎及其他相关肺炎的中文对话，共有来自 935 名患者和 352 名医生的话语 9 494 条，平均每个话语包含 42.8 个词。候选实体集合使用 TF-IDF（term frequency-inverse document frequency）技术从数据集中提取。每段对话由 3 个部分组成：对患者病情和病史的描述、患者与医生之间的对话、由医生提供的诊断和治疗建议。本文将该数据集划分为训练集（870 段对话）、验证集（109 段对话）和测试集（109 段对话）。

MedicaKnowledgeGraph 是一个以垂直型医药网站为数据来源的知识图谱，以疾病为核心，包含 7 类规模为 4.4 万的知识实体，11 类规模约 30 万实体关系的知识图谱。

3.1.2　数据预处理

数据预处理操作主要有 MedicaKnowledgeGraph 知识图谱预处理、医疗知识图谱子图获取、历史对话和话语角色获取。

首先，利用 MedicaKnowledgeGraph 知识图谱生成以疾病实体为中心的三元组集合，其中每个三元组由两个实体和它们之间的关系组成，例如肩关节、symptom（症状）、骨质疏松。

然后，根据两个数据集的划分情况，利用医生和患者的历史对话获取每段对话中医生和患者的全部话语，根据三元组集合获取医疗知识图谱子图信息并保存在相应的文件中，其中包含实体、关系以及对应的图注意力机制矩阵。

最后，根据两个数据集的划分情况，获取每段对话对应的全部历史对话以及每个话语对应的角色，并保存在相应的文件中。

3.1.3　基线模型

由于本文方法旨在根据历史对话和医疗知识图谱生成预问诊回复，因此选择目前最优且相关的基于知识库的回复生成模型进行对比，并从相关文献中直接抽取实验结果。

Retrieval：基于检索的回复生成方法，使用 Lucene 搭建检索引擎。它使用最相关的实体从知识库中检索最相关的回复。

Seq2Seq[14]：一种经典的基于注意力机制的序列到序列模型，由编码器和解码器组成，其中编码器负责编码历史对话，解码器负责解码得到的最终回复。

HRED[15]：在 Seq2Seq 的基础上，采用分层循环神经网络（RNN）对上下文进行建模。通过分层堆叠两个 RNN 扩展了传统的 RNN 编码器，一个在单词级别，一个在话语级别。

DialoGPT[16]：DialoGPT 模型在 GPT-2 基础上进行扩展，也为自回归语言模型，适合用于处理生成式任务，使用了多层 Transformer 解码器作为模型架构。但不同于 GPT-2 模型，DialoGPT 模型的训练使用了从 Reddit 讨论链中提取出的大规模对话数据，是 GPT-2 模型的一个变体，实现了最大化互信息评分方程。

Transformer[17]：Transformer 模型抛弃了传统的 RNN 和卷积神经网络（CNN）结构，设计了一种注意力机制，它由且仅由自注意力机制和前馈神经网络组成，使用堆叠的 Encoder-Decoder 结构搭建。

BERT-GPT[18]：Transformer 模型的一个预训练方法，其中 BERT 用于预训练 Transformer 编码器部分，GPT 用于预训练 Transformer 模型解码器部分。

TAPT[19]：任务自适应预训练技术。先对任务相关的无标注语料进行预训练，然后再对特定任务进行微调。

BERT-GPT-TAPT[13]：在 BERT-GPT 模型的基础之上使用了任务自适应预训练技术。

MCG-HE：基于历史对话和实体预测的医疗对话生成模型，主要包含实体预测模块和回复生成模块，实体预测模块使用 BERT 预测与历史对话最相关的关键指导实体，回复生成模块将关键指导实体以及历史对话作为 GPT-2 模型的输入实现回复生成。

3.1.4　实验参数

本文模型 MCG-KE 使用 PyTorch 深度学习框架实现，其中子图编码模块设置：batch size 为 4，学习率为 6.0×10^{-5}，epoch 为 3，所使用的中文编码器是 BertTokenizerFast；回复生成模块使用 $M=12$ 层 GPT-2 预训练模型，设置 batch size 为 1，学习率为 1.0×10^{-5}，epoch 为 3，模型设置每句生成对话的最长长度为 50 个词，基线模型部分参数设置见表 2。

3.1.5　评估指标

采用自动评估和人工评估两种方法评估 MCG-KE 方法。

自动评估指标：又称为客观评价指标。本文参考对话系统采用 BLEU（bilingual evaluation understudy）和 Distinct 两个指标。

BLEU 指标评估生成回复与真实回复的词重叠程度，计算公式为

$$\mathrm{BLEU}=\mathrm{BP}\cdot\exp\left(\sum_{n=1}^{N}W_n\log P_n\right) \tag{8}$$

表 2　基线模型部分参数设置

Tab. 2　Parameter configurations of baseline model

模型	参数设置
BERT-GPT	12 层，hidden states 为 768，学习率为 1×10^{-4}，最大序列长度为 400
GPT-2	12 层，context size 为 300，embedding size 为 768，多头注意力机制中多头 12 个，学习率为 1.5×10^{-4}，batch size 为 8
Transformer	编码器 6 层，解码器 6 层，隐层 512 维，多头 8 个
DialoGPT	12 层，多头 12 个，学习率为 1.5×10^{-4}，batch size 为 8
Seq2Seq	4 层 LSTM，词向量维度为 1 000，学习率为 0.7
HRED	最大序列长度为 80，batch size 为 8，优化算法为 Adam
MCG-HE	实体预测模块使用 $N=12$ 层 BERT 预训练模型，设置字向量维度为 768，batch size 为 8，学习率为 1×10^{-5}，epoch 为 500；回复生成模块使用 $M=12$ 层 GPT-2 预训练模型，设置 batch size 为 15，学习率为 1.5×10^{-4}，epoch 设置为 50，模型设置每句生成对话的最长长度为 50 个词

$$\mathrm{BP}=\begin{cases}1,\text{if } c>r\\ \mathrm{e}^{1-r/c},\text{if } c\leqslant r\end{cases} \tag{9}$$

其中：c 为生成的回复句子长度；r 为参考回复句子长度；W_n 指 n-gram 的权重，n-garm 是包含 N 个词的连续词片段，一般设为均匀权重，即对于任意 n 都有 $W_n=1/N$；P_n 是指 n-garm 的精度，计算公式为

$$P_n=\frac{\sum c\in\text{candidates}\sum n\text{-gram}\in c\ \mathrm{Count}_{\mathrm{clip}}(n\text{-gram})}{\sum c'\in\text{candidates}\sum n\text{-gram}'\in c'\ \mathrm{Count}_{\mathrm{clip}}(n\text{-gram}')} \tag{10}$$

其中：candidates 是参考回复，$\mathrm{Count}_{\mathrm{clip}}(n\text{-gram})$ 表示某一个 n-gram 在生成回复中的个数，$\mathrm{Count}(n\text{-gram}')$ 表示 n-gram′在 candidates 中的个数。

Distinct 指标评估生成回复在词级别的多样性。Distinct-1 为所有生成回复中不同的 1-gram 占所有 1-gram 的比例。Distinct-2 为所有生成结果中不同 2-gram 占所有 2-gram 的比例。Distinct 指标取值越高，表示生成的结果中含有越多不同的 n 元组，说明结果的多样性更好。Distinct-n 计算公式为

$$\text{Distinct-}n=\frac{\mathrm{Count}_{\mathrm{unique}}(n\text{-gram})}{\mathrm{Count}(n\text{-gram})} \tag{11}$$

其中：$\mathrm{Count}_{\mathrm{unique}}(n\text{-gram})$ 表示回复中不重复的 n-gram 词语数量，$\mathrm{Count}(n\text{-gram})$ 表示回复中 n-gram 词语的总数量。

人工评估指标：又称为主观评价指标，是对话系统中最重要的评估指标，能表示对话系统模拟人工的有效性。本文使用了 correctness、relevance、informativeness、doctor-likeness 这 4 个指标评估生成回复的质量。correctness 是指这个回复在临床上的正确程度，relevance 是指回复与历史对话的相关程度，informativeness 是指在回复中提供了医疗信息和建议的信息量，doctor-likeness 是指生成回复与真实医生回复的相似程度。本文邀请 4 位医学相关专业的专家根据相关医学知识对生成回复进行评估。他们对所有生成回复的 4 个指标从 1～5 进行打分，分数越高代表效果越好。计算所有生成回复在每个指标上的平均分数，并作为最终结果。

3.2　实验结果与分析

3.2.1　自动评估

在 MedDG 数据集和 CovidDialog-Chinese 数据集上的自动评估实验结果分别见表 3 和表 4，其中-e 表示模型不使用实体预测模块。本文选取已发表论文中所汇报的在两个数据集上的先进方法进行对比，这几种对比模型的实验结果直接来自相关参考文献。为了让实验结果更加清晰、直观，本文把所有自动评估实验结果都归一化为[0，100]之间，并且最优结果以粗体表示。

表 3　在 MedDG 数据集上的自动评估指标对比

Tab. 3　Comparison of automated assessment indicators on MedDG datasets

模型	BLEU-1	BLEU-4	Distinct-1	Distinct-2
Retrieval	23.08	12.58	0.62	9.98
Seq2Seq	26.12	14.21	0.88	4.77
HRED	31.56	17.28	1.07	8.43
DialoGPT	34.57	18.09	0.50	9.92
MCG-HE(-e)	14.15	0.53	1.70	9.48
MCG-HE	17.09	1.02	**7.00**	**25.45**
MCG-KE	**75.62**	**28.52**	1.21	2.49

表 4　在 CovidDialog-Chinese 数据集上的自动评估指标对比

Tab. 4　Comparison of automated assessment indicators on CovidDialog-Chinese datasets

模型	BLEU-2	BLEU-4	Distinct-1	Distinct-2
Transformer	5.70	4.00	5.50	29.00
BERT-GPT	4.60	2.80	7.90	39.50
BERT-GPT-TAPT	5.10	2.60	9.10	**39.70**
MCG-HE(-e)	3.97	2.63	11.10	32.70
MCG-HE	7.61	3.56	**12.97**	37.00
MCG-KE	**42.24**	**12.22**	1.80	3.60

从表 3 可以看出，在 MedDG 数据集上，MCG-KE 模型在 BLEU 指标上优于所有的基线模型，而在 Distinct 指标上，MCG-KE 没有取得很好的表现，不如 MCG-HE 模型。猜测可能的原因是 Distinct 指标不适用于此任务[20]，该指标用于度量生成回复的多样性，多样性越高并不代表生成回复的质量就越好，

由于医疗对话生成属于任务型对话，模型应该根据患者的需求给出符合医学原理的正确回复，因此更加关注生成的回复是否准确，而不是多样。

从表 4 可以看出，在 CovidDialog-Chinese 数据集上，MCG-KE 模型在 BLEU 指标上优于所有的基线模型，同时 MCG-HE 模型也比基线模型效果好；但如果不使用实体预测模块，指标均有下降，说明实体的加入对生成回复有积极作用。

从以上两个数据集的实验结果可以看出，在 BLEU 指标上 MCG-KE 模型取得了比较好的结果，说明 MCG-KE 模型生成的回复更加接近于真实回复，医疗知识图谱和上下文知识嵌入能够有效指导模型生成更加准确、合适的回复。

3.2.2　人工评估

在 MedDG 数据集和 CovidDialog-Chinese 数据集上的人工评估实验结果见表 5 和表 6，其中最优结果用粗体表示。与自动评估不同的是，本文选择在两个数据集上与相关学者的实验结果进行对比。

表 5　在 MedDG 数据集上的人工评估指标对比

Tab. 5　Comparison of manual assessment indicators on MedDG datasets

模型	correctness	relevance	informativeness	doctor-likeness
GPT-2	2.80	2.85	2.86	3.15
HRED	2.70	3.02	2.95	3.33
MCG-HE	3.18	3.36	2.95	3.30
MCG-KE	**3.31**	**3.46**	**3.29**	**3.41**

表 6　在 CovidDialog-Chinese 数据集上的人工评估指标对比

Tab. 6　Comparison of manual assessment indicators on CovidDialog-Chinese datasets

模型	correctness	relevance	informativeness	doctor-likeness
Transformer	1.94	2.09	2.03	2.61
BERT-GPT	2.15	2.70	2.32	3.02
TAPT	2.27	2.68	2.42	3.11
MCG-HE	2.79	3.05	2.80	3.29
MCG-KE	**3.02**	**3.11**	**3.04**	**3.36**

从表 5 和表 6 可以看出，在两个数据集上，MCG-KE 模型在 4 个人工评估指标上都取得了最优结果，相比于基线模型有明显提升。同时，MCG-HE 模型比基线模型效果好，再次证明了实体的引入有助于生成回复。另外，从实验结果可以看出，MCG-KE 模型比 MCG-HE 模型效果好，模型在引入了医疗知识图谱之后，模型生成的回复被认为更加正确，与上下文更加相关，信息量更加丰富，由此说明医疗知识图谱含有丰富的实体以及关系信息，能够指导模型生成更加接近于人类医生的回复。

4　结　语

本文提出了知识嵌入的医疗对话生成模型 MCG-KE，该模型通过历史对话预测最相关的上下文嵌入实体，利用图注意力机制表示实体之间的关系，使用串行图编码技术编码医疗知识图谱子图，将医疗知识图谱集成到基于知识的医疗对话系统当中，指导模型生成更加接近于人类医生的回复，提升系统的推理鲁棒性。在实验部分，通过在两个数据集上的实验结果验证了本文方法的有效性。知识图谱在不同实体之间的复杂关系可以被简洁地编码，并且不会导致知识正确性、一致性和信息性等关键指标下降。

参考文献：

[1] 陈建平，蔡俊，鞠沁怡，等. 江苏省优质医疗资源配置现状分析与思考[J]. 江苏卫生事业管理，2023，34(1)：1–3.

[2] LIU W，TANG J，QIN J，et al. MedDG：a large-scale medical consultation dataset for building medical dialogue system[EB/OL]. [2023–01–05]. https://arxiv.org/abs/2010.07497v1.

[3] 刘源. 基于知识图谱的医疗问答系统[D]. 成都：电子科技大学，2021.

[4] CHAUDHURI D，RONY M R A H，LEHMANN J. Grounding dialogue systems via knowledge graph aware decoding with pre-trained transformers[C]//The Semantic Web：18th International Conference，ESWC 2021. Berlin：Springer International Publishing，2021：323–339.

[5] LIU Z，NIU Z Y，WU H，et al. Knowledge aware conversation generation with explainable reasoning over augmented graphs[C]//EMNLP. Proceedings of the 2019

Conference on Empirical Methods in Natural Language Processingand the 9th International Joint Conference on Natural Language Processing. New Haven：EMNLP，2019：1782–1792.

[6] LIU S，ZHAO X，LI B，et al. A three-stage learning framework for low-resource knowledge-grounded dialogue generation[EB/OL]. [2023–01–05]. https://doi.org/10.48550/arXiv.2109.04096.

[7] LIN X，HE X，CHEN Q，et al. Enhancing dialogue symptom diagnosis with global attention and symptom graph [C]//ACL. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing（EMNLP-IJCNLP）. Hong Kong：ACL，2019：5033–5042.

[8] LIN S，ZHOU P，LIANG X，et al. Graph-evolving meta-learning for low-resource medical dialogue generation [EB/OL]. [2023–01–05]. https://arxiv.org/abs/2012.11988.

[9] XU L，ZHOU Q，GONG K，et al. End-to-end knowledge-routed relational dialogue system for automatic diagnosis[C]//Proceedings of the 33th AAAI Conference on Artificial Intelligence. Hawaii：Association for the Advancement of Artificial Intelligence，2019：7346–7353.

[10] 章毅，郭泉，张海仙，等. 一种基于语言模型和实体匹配的医疗问答系统构建方法：112667799B[P]. 2021–06–01.

[11] 穆天杨，李宝安，游新冬，等. 一种融合医疗知识图谱的端到端对话系统[J]. 北京信息科技大学学报（自然科学版），2021，36（6）：14–18.

[12] JACOB D，CHANG M，LEE K，et al. Bert：pretraining of deep bidirectional transformers for language understanding[EB/OL]. [2023–01–05]. https://arxiv.org/pdf/1810.04805.pdf.

[13] ZHOU M，LI Z，TAN B，et al. On the generation of medical dialogs for COVID-19[C]//ACL. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing（Volume 2：Short Papers）. Virtual Event：Association for Computational Linguistics，2021：886–896.

[14] ILYA S，ORIOL V，LE QUOC V. Sequence to sequence learning with neural networks[C]//Advances in neural information processing systems. Canada：Neural Information Processing Systems Foundation，2014：3104–3112.

[15] IULIAN S，ALESSANDRO S，YOSHUA B，et al. Building end-to-end dialogue systems using generative hierarchical neural network models[C]//AAAI. Proceedings of the 30th AAAI Conference on Artificial Intelligence. Phoenix：Association for the Advancement of Artificial Intelligence，2016：1–8.

[16] ZHANG Y，SUN S，GALLEY M，et al. Dialogpt：large-scale generative pre-training for conversational response generation[EB/OL]. [2023–01–05]. https://arxiv.org/abs/1911.00536.

[17] ASHISH V，NOAM S，NIKI P，et al. Attention is all you need[C]//Advances in Neural Information Processing Systems. Long Beach：Neural Information Processing Systems Foundation，2017：5998–6008.

[18] WU Q，LI L，ZHOU H，et al. Importance-aware learning for neural headline editing[C]//AAAI. Proceedings of the 34th AAAI Conference on Artificial Intelligence. New York：Association for the Advancement of Artificial Intelligence，2020：9282–9289.

[19] SUCHIN G，ANA M，SWABHA S，et al. Don’t stop pretraining：adapt language models to domains and tasks [EB/OL]. [2023–01–05]. https://arxiv.org/abs/2004.10964v3.

[20] LIU C W，LOWE R，SERBAN I V，et al. How not to evaluate your dialogue system：an empirical study of unsupervised evaluation metrics for dialogue response generation[EB/OL]. [2023–01–05]. https://doi.org/10.48550/arXiv.1603.08023.

责任编辑：郎婧